A Common Thermal Network Problem Encountered in Heat Transfer Analysis of Spacecraft

Mark Milman and Walt Petrick

Jet Propulsion Laboratory

A Common Thermal Network Problem Encountered in Heat Transfer Analysis of Spacecraft

0. Introduction. A widely used discretization method for modeling thermal systems is the thermal network approach. The network approach is derived from the energy balance equations and is equivalent to a particular finite difference discretization of the underlying heat transfer equation. The thermal network is defined by a set of set of nodes and conductances, and is analogous to an electrical network. Thus there is a correspondence between potential, flow, resistance, capacitance, etc., between these networks, and basic laws such as Ohm's Law and Kirchoff's Laws can then be applied to balancing the network [9].

Although there is the underlying connection with the heat transfer equation defined over a spatial domain, the approach is also used to develop models of complex systems for which a node may represent a particular isothermal component of the system which does not necessarily conform to a regular discretization of the domain. This is a commonly encountered situation in spacecraft thermal analysis problems. For example, an engineer may use a single node to model an electronics box or a cryogenically cooled component of a spacecraft. Irrespective of these considerations, the domain is divided into a number of elements, or subvolumes, and a node is typically placed at the center of the element. The temperature of the node is the average temperature over the subvolume. In addition to temperature, a node also has a capacitance which represents the thermal mass of the node that dictates how quickly the node can change temperature. The capacitance is computed from the properties of the material comprising the subvolume attached to the node, and arises in the analysis of the transient characteristics of the system. Our interest here is primarily in the steady state problem so the nodes we consider will have essentially a zero time constant (these are sometimes referred to as arithmetic nodes).

The conductors describe the energy transport between nodes. They have the form of either a conduction conductor, a convection conductor, or a radiation conductor. The underlying physical mechanisms for each of these modes of energy transfer is different. Within the context of the resulting mathematical model, conduction and convection conductors transfer energy as the difference between the temperatures between connected nodes, while the radiator conductors are nonlinear and involve the difference in the fourth power of the nodal temperatures. Conductors normally have the same value in both directions between nodes; however, one way nodes are often used to model fluid flow.

The resulting transient equations have the form

$$C_i \frac{dT_i}{dt} = Q_i + \sum_{j=1}^{N} C_{ij}(T_j - T_i) + \sum_{j=1}^{N} R_{ij}(T_j^4 - T_i^4); \quad i = 1, ..., N \tag{0.1}$$

where T_i denotes the temperature at node i, C_i is the heat capacitance of node i, C_{ij} are the conduction coefficients, R_{ij} are the radiation coefficients, and the Q_i are heat sources. The associated steady state equation we study in this note is

$$Q_i + \sum_{j=1}^{N} C_{ij}(T_j - T_i) + \sum_{j=1}^{N} R_{ij}(T_j^4 - T_i^4) = 0; \quad i = 1, ..., N \tag{0.2}$$

The conduction coefficients are computed as a function of the thermal conductivity of the material, the cross sectional area through which the heat flows, and the length between nodes. For a rectangular discretization Fourier's Law yields

$$C_{ij} = \frac{kA}{L} \tag{0.3}$$

where k denotes the thermal conductivity of the material, A is the cross sectional area, and L is the length between nodes i and j. For convection conductors,

$$C_{ij} = hA,$$

where h is the thermal convective conductance and A is the nodal surface area in contact with the fluid. The thermal conductivity is in general temperature dependent, so that the C matrix above is also. The radiation interchange matrix R is a function of the the surface geometries, the orientation of the elements of the system with respect to one another (view factors), their radiative properties, and temperature. Typically, $C_{ij} = C_{ji}$, and $R_{ij} = R_{ji}$, although when modeling fluid flow, the use of one way nodes causes asymmetry in the matrix C. When temperatures are not expected to deviate greatly from a nominal value, the assumption that the coefficients of C and R are temperature independent, i.e. constant, is often made. At cryogenic temperatures, where small variations in temperature can lead to significant changes in the thermal conductivity of materials this is not a valid assumption.

There exist a number of commercial codes that have enjoyed considerable success solving the steady state network problem (0.2). Solution techniques often rely on exploiting specific structure of the network. Nonlinear Jacobi or Gauss Seidel iteration schemes which employ an analytic solution of individual the scalar quartic equations in (0.2) are used, as well as variations of Newton Raphson methods that, take advantage of the sparsity inherent in most network problems, and various other successive substitution methods used in conjunction with acceleration techniques [5, 9, 13, 20].

Although (0.2) is a very commonly occuring problem, there does not appear any general proof of the existence of positive solutions, even for the constant coefficient problem. For mildly non-linear problems where the coefficient matrix R has small norm, the Newton Kantorovich theorem produces a solution under appropriate hypotheses. However, in many spacecraft models this condition cannot be assumed, as radiation is often the dominant mode of heat transfer. And although (0.2?) has a relatively rich structure, it does not fit into the several classes of nonlinear equations for which general convergence results are available, e.g. M functions, or diagonally dominant functions ["14, 15]

In this paper a homotopy approach is first used to prove existence and uniqueness of positive solutions to (0.2) in the constant coefficient case with a mild restriction on the conduction and radiation coefficients and heat loads. This result enables the construction of a globally convergent algorithm. It also suggests other globally convergent schemes, and explains the relative success of many algorithms for solving these thermal network problems under similar conditions. We also show that the hypotheses of this problem are easily generalized to include solutions to other discretized boundary value problems. The constant coefficient solution is then used to initialize a second homotopy for proving existence of positive solutions to (0.2) for temperature dependent coefficients under mild growth restrictions. We note that highly nonlinear problems that occur in chemical processes may not even possess solutions [20].

Homotopy methods have been previously employed for solving a number of other engineering problems. Typically they perform somewhat slower than other methods, but for difficult problems they offer a robust solution approach [1, 2, 11, 17]. A small sample problem compares the performance of the Newton method with stepsize control, the nonlinear Gauss Seidel iteration, and a continuation method.

1. Preliminaries. The steady state equation of heat transfer we study in this note has the

form

$$Q_i + \sum_{j=1}^{N} C_{ij}\,(T_j - T_i) + \sum_{j=1}^{N} R_{ij}(T_j^4 - T_i^4) = 0, \quad i = 1,\ldots,N \tag{1.1}$$

Here T_i denotes the temperature at node i, C_{ij} are the conduction coefficients, R_{ij} are the radiation coefficients, and the Q_i are heat sources. Initially we will assume that the conduction and radiation coefficients are independent of temperature.

We begin the analysis by introducing the $N \times N$ matrices $\tilde{C}$ and $\tilde{R}$ defined

$$C_{ij} = \begin{cases} C_{ij} & \text{if } i \neq j \\ -\sum_{j=1}^{N} C_{ij} & \text{if } i = j, \end{cases} \tag{1.2a}$$

and

$$R_{ij} = \begin{cases} R_{ij} & \text{if } i \neq j \\ -\sum_{j=1}^{N} R_{ij} & \text{if } i = j. \end{cases} \tag{1.2b}$$

Let $\tilde{Q} = [Q_1, \ldots, Q_N]^T$, and $T = [T_1, \ldots, T_N]^T$, and define $\tilde{D} : R^N \to R^N$ by $D(T) = [T_1^4, \ldots, T_N^4]^T$. With these notations the system of scalar equations (1.1) can be written as

$$\tilde{C}T + \tilde{R}\tilde{D}(T) + \tilde{Q} = 0. \tag{1.3}$$

Next we assume without loss of generality that $\tilde{C}$, $\tilde{R}$, $\tilde{Q}$ and T have been partitioned to have the form

$$\tilde{C} = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}, \qquad \tilde{R} = \begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix}, \qquad T = \begin{pmatrix} T_1 \\ T_2 \end{pmatrix}, \qquad \tilde{Q} = \begin{pmatrix} Q_1 \\ Q_2 \end{pmatrix} \tag{1.4}$$

where T_2 denotes the fixed and known boundary nodes. Using these values, the equation for the unknown nodes becomes

$$C_{11}T_1 + R_{11}D_1(T_1) + [Q] + C_{12}T_2 + R_{12}D_2(T_2) = 0. \tag{1.5}$$

Here $\tilde{D}$ has been partitioned as $\mathrm{diag}(D_1\, D_2)$.

In the sequel we shall write $C = C_{11}$, $R = R_{11}$, $D = D_1$, $x = T_1$, and $Q = Q_1 + C_{12}T_2 + R_{12}D_2(T_2)$, and study the equation

$$F(x) = 0; \quad F(x) = Cx + RD(x) + Q \quad \text{with} \quad dim(x) = n. \tag{1.6}$$

The entries of C shall be denoted c_{ij} and the entries of the matrix R will be denoted r_{ij}. Note that by construction both C and R are diagonally dominant matrices, although not necessarily strictly diagonally dominant. (A matrix $A = (a_{ij})$ is diagonally dominant if $\sum_{i \neq j} |a_{ij}| \leq |a_{ii}|$. Strict diagonal dominance holds when strict inequality holds for all i.) An important characteristic of the network equation that will be exploited in the sequel is the assumption that the matrix $C + R$ is *irreducible*.

An $n \times n$ matrix $A = (a_{ij})$ is irreducible if its directed graph $G(A)$ is strongly connected, that is, for any pair of indices i, j there is a sequence of nonzero entries of the form $(a_{ir}, a_{rs}, a_{st}, \ldots, a_{uj})$ [3]. This condition has a simple physical interpretation for the sum of the conduction and radiation interchange matrices $C + R$: Given a pair of nodes i, j, there is a sequence of nodes $r, s, t, \ldots, u$ connecting i and j such that the interchange factors between each successive pair is nonzero. Thus each node can transfer energy to ally other node through a sequence of nodes connected by a combination of conductors or radiators. Under these reasonable physical assumptions, the

hypothesis C I R is irreducible is satisfied. Furthermore, if at least one node is connected to a boundary node with nonzero interchange factor, then C | R has strict diagonal dominance in the row corresponding to this node. In this case C -| R is an *irreducibly dominant* matrix [16]. An irreducibly dominant matrix is invertible [1 6].

2. M ain Results. The existence results established in this section u s(: a simple continuation idea. Before this idea can be applied h owever, we first need to show that t he function F is coercive for fixed Q, i.e., the energy balance equation cannot be sat isfied by arbitrarily large t e mperature values.

Lemma 2.1. Let C and R be diagonally dominant symmetric nonegative definite matrices with C -| R irreducibly dominant. Then $\lim_{|x|- \to \infty} |Cx$ -| $RD(x)| = \infty$.

Proof. First suppose that R is invertible. Let $|x_N| - \to \infty$ and define $v_N = x_N/|x_N|_\infty$. Then

$$|Cx_N \text{ -| } RD(x_N)|_\infty = |x_N|_\infty^4 |C\frac{v_N}{|x_N|^3} \text{ -| } Rv_N|_\infty.$$

Now,

$$|C\frac{v_N}{|x_N|^3}|_\infty - \to 0 \quad \text{as} \quad N - \to \infty,$$

and there exists $\rho > 0$ such that $|Rv_N| \geq \rho$ for all N since R is invertible. 1 hence, $|Cx_N$ | $RD(x_N)| - \to \infty$.

So now assume that R is not invertible. (We may assume that R is nonzero, since if it were this would require C to be invertible, and hence coerc ive.) In this case we may write, after a permutation of coordinates, R as a block diagonal matrix

$$R = diag(R_1, ..., R_k, O_{n_k \times n_k}),$$

where R_i is an $n_i \times n_i$ irreducible matrix for each i and $O_{n_k \times n_k}$ is the $n_k \times n_k$ zero matrix. For $j = 1, ..., k$ let $e^j = [1 \cdots 1]^T \in R^{n_j}$ and let E^j denote the subspace generated by e^j. Note that if R_j is not invertible, then $N(R_j) = E^j$, since $R_j e^j = 0$ and R_j becomes invertible by adding a rank one perturbation corresponding to a positive (negative) entry on the diagonal (hence $dim(N(R_j)) \leq 1$. Let $S = N(R)$. Then $S \subset \oplus E^j$.

Suppose there is a sequence x_N with $|x_N| - \to \infty$ such that $|Cx_N$ -| $RD(x_N)|$ remains bounded. Let Π denote the orthogonal projection onto S, and let $v_N = x_N/|x_N|$. Then

$$|RD(v_N)|^2 \geq o_{min} |(I - \Pi)D(v_N)|^2,$$

where o_{min} denotes the sm allest nonzero eigenvalue of Π'. Now

$$|(I - \Pi)D(v_N)|^2 \quad 1 - |\Pi D(v_N)|^2,$$

so that

$$|RD(v_N)| \geq o_{min} \sqrt{1 - |\Pi D(v_N)|^2}$$

Let $\rho_N = |\Pi D(v_N)|^2$. Since $|x_N|^4 RD(v_N)$ and $|x_N||Cv_N|$ are of the same magnitude, it follows that $\rho_N - \to 0$. And hence, $D(v_N) - \to S$. But since $S \subset \oplus E^j$, we have also that $v_N - \to S$.

Now let $s \in S$. Since R is symmetric,

$$< RD(x_N), \text{ s } > 0.$$

However, since $C + R$ is invertible, $N(C) \cap S = 0$. Hence, $R(C)^{\perp} \cap S = 0$ and consequently

$$\min_{|v|=1} | < Cv, s > | \geq m > 0,$$

for some m. 'I'1111s,

$$\lim_{N \to \infty} | < Cx_N, s > | = \lim_{N \to \infty} |x_N| < Cv_N, s > | = \infty$$

and hence, $|RD(x_N) + Cx_N|) \infty$. ///

The method of proof is based on the following simple homotopy idea: Let x_0 denote an initial estimate of the solution, and define Q_0 by

$$Q_0 = Cx_0 + RD(x_0).$$

For $\lambda \in [0, 1]$ consider the function

$$H(x, \lambda) = Cx + RD(x) + Q_0 + \lambda(Q - Q(). \tag{2.1}$$

Note that $H(x_0, 0) = 0$, and the original problem is to solve $H(x,1) = 0$. Showing that solutions to $H(x, \lambda) = 0$ can be continued from $\lambda = 0$ to $\lambda = 1$ involves properties of the differential of H. For this we note that

$$H'(x) \cdot h \to Ch + RD_x h, \text{ where } D_x(h) [4x_1^3 h_1, \ldots, 4x_n^3 h_n]^T. \tag{2.2}$$

Hence, the differential D_x is the diagonal matrix with i^{th} entry $4x_i^3$. 'I'11(! theorem below Snows that this curve can always be continued.

Theorem 2.2, Suppose that C and R are symmetric nonpositive matrices with $C + R$ irreducibly dominant. Then if Q is nonzero with $Q \geq 0$, (1.6) has a unique solution x^* with $x^* > 0$.

Proof." Let $x_0 = \rho[1 \ldots 1]^T$ with $\rho > 0$ and observe that $Q_0 = Cx_0 - RD(x_0) \geq 0$. Q, $Q_0 \geq 0$ implies that $Q_0 + \lambda(Q - Q(1) \geq 0$ for all $\lambda \in [0,1]$ by convexity. Note that by (1.2) and (1.4) both C and R are diagonally dominant matrices with negative values on the diagonal. Note also that so long as $x > 0$, $\partial H/\partial x$ is invertible, since the matrix $[C + RD_x]^T$ is easily shown to be irreducibly dominant. By the implicit function theorem we can locally solve (2.1) in a neighborhood of $\lambda > 0$ to produce a smooth solution curve. This curve can be continued so long as $\partial H/\partial x$ remains invertible. And this happens as long as $x > 0$.

To show that solutions stay bounded away from zero, suppose there exist values of $\lambda < 1$ such that some component of the solution $x(\lambda)$ is zero. Let λ^* denote the infimum of all such λ. By continuity $x(\lambda^*) \geq 0$, i.e. $x(\lambda^*)$ has no negative components. We will show that this is impossible. Suppose one or more components of $x(\lambda^*)$ is zero. It is clear that x is not identically zero since $Q(0) + \lambda^*(Q - Q(0)) > 0$. Assume that $x_j = 0$ but $x_k > 0$. Since $C + R$ is irreducible, there exists a sequence $\{c_{jj_1} + r_{jj_1}, c_{j_1 j_2} + r_{j_1, j_2}, \ldots, c_{j_n k} + r_{j_n k}\}$, with each entry nonzero. It is easy to deduce that $x_j = 0$ implies $x_{j_1} = 0$, which in turn implies $x_{j_2} = 0$, etc. Hence, $x_k = 0$, which is a contradiction, and solutions of (2.1) with $x_0 > 0$ must remain positive.

Now we can assert that the Set

$$S \{(x, \lambda) \in (0,1) \times R^n : H(x, \lambda) = 0\}$$

is a one dimensional manifold. The connected components of S are either diffeomorphic t o a circle or the real line. The solution curve $x(\lambda)$ must be diffeomorphic to the real line since $\partial H/\partial x \quad F'(x)$ remains invertible, and thereby unique solutions for (2. 1) exist in every neighborhood. The only way $x(\lambda)$ could not be continued to $\lambda \quad 1$ is if for some λ^*, $lim_{\lambda\text{-}\rightarrow\lambda^*}|x(\lambda)| \quad$ cm. By Lemma 2.1, $|Cx \mid RD(x)|$ -, ∞ as $|x|$ -, c-o. This then implies that solutions to (2.1) ccase to exist in an entire neighborhood of λ^* since Q and Q_0 are finite. Thus $x(\lambda)$ can be continued.

To show uniqueness, suppose that x and y are two positive solutions. Then

$$\begin{aligned} 0 \quad & F(y) \ - F(x) \\ & C(y \ - \ x) \mid R[D(y) \ - \ D(x)] \end{aligned}$$

But,

$$\begin{aligned} D(y) \ - \ D(x) \quad & \begin{pmatrix} (x_1^2 \mid y_1^2 \ )(x_1 \mid y_1 \ )(y_1 \ x_1) \\ \\ (x_n^2 \ \mathrm{I} \ y_n^2)(x_n \ \mathrm{I} \ y_n)(y_n \ x_n) \end{pmatrix} \\ & : \ D_z(y \ - \ x), \end{aligned}$$

where

$$z : \quad \begin{pmatrix} (x_1^2 \mid y_1^2)(x_1 \mid y_1) \\ \vdots \\ (x_n^2 \mid y_n^2)(x_n \mid y_n) \end{pmatrix}$$

Thus

$$C(y \ x) \mid RD_z \ (y \ - \ x) \quad 0.$$

But $C \mid RD_z$ is invertible for $z > \ ()$. Thus $y \quad x.///$

The proof of Theorem 2.2 relied on properties of the matrices C, R, and $D(x)$ to establish the invertibility of F', the coercivity of F, and the positivity of solutions along the homotopy path. These intermediate results form the crux of the proof of the theorem. If instead we *assume* these properties, then the argument of the theorem is transparent. With this idea in mind, consider the equation

$$Cx \mid \phi(x) \mid Q : \ 0, \quad \text{with} \quad Q \geq 0, \quad Q \ / \ 0. \tag{2.10}$$

Suppose ϕ is continuously differentiable and the following four conditions hold:

(i) There exists $x_0 > 0$ and $Q_0 \geq 0$ with $Q_0 \ / \ 0$ such that $\mathrm{Cx_0} \mid \phi(x_0) \mid Q_0 : \ 0$,

(ii) $C \mid \phi'(x)$ is invertible for all $x \geq 0$,

(iii) $|Cx \mid \phi(x)| \text{ - } \rightarrow \infty$ as $|x| \text{ - } \rightarrow \infty$,

(iv) If x^* solves

$$Cx \mid \phi(x) \mid Q^* : \ 0, \quad \text{with} \quad Q^* \geq 0, \quad Q^* \ / \ 0,$$

then $x^* > 0$.

The method of proof in Theorem 2.2 shows that (2.10) has a positive solution.

Example 2.3. As an example of how this set of hypotheses arises in another setting, consider the discretization of the two point boundary value problem:

$$u''(x) \quad g(x, u, u'), \quad a < x < b, \tag{2,"11 (,}$$

$$u(a) \quad \alpha, \quad u(b) \quad \beta. \tag{2.11b}$$

Here it is assumed that g is continuously differentiable on

$$\Omega = \{(x,u,u') : \quad x \in [a,b], u \geq 0, u' \in R\},$$

with $g_u \geq 0$, $|g_{u'}| \leq M < +\infty$, $g(x,0,0) \leq 0$, and $\alpha, \beta \geq 0$. Define the uniform grid $\{x_j\}_{j=0}^{j=N}$ where

$$x_0 = a, \quad x_{j+1} = x_j + h, \quad j = 0,1,...,N+1 \quad \text{with} \quad h = (b-a)/(N+1)$$

The finite difference approximation to (2.11) has the form

$$Cu + \phi(u) - Q = 0, \tag{2.12}$$

where C is the $N \times N$ tridiagonal matrix $C = tri(1,-2,1)$ (indicating -2's on the diagonal, and 1's on the upper and lower diagonals), and $\phi = (\phi_i)$, $i = 1,...,N$, with

$$\phi_i(u) = -h^2 g(x_i, u_i, \frac{u_{i+1} - u_{i-1}}{2h}), \tag{2.13}$$

and $Q \in R^N$, $Q = [\alpha, 0, \cdots, \beta]$.

We will now outline how (i)-(iv) are satisfied for this model for sufficiently small h. (Specifically, $h < 2/M$). First introduce $\psi = (\psi_i)$,

$$\psi_i(u) = \phi_i(u) + h^2 g(x_i, 0, 0), \quad i = 1, \quad ..., N, \tag{2.14}$$

and note that $Cu + \phi(u) - Q = 0$ and $Cu + \psi(u) - \tilde{Q} = 0$ are equivalent with $\tilde{Q}$ defined as $\tilde{Q} = Q + [g(x_1, 0, 0), \ldots, g(x_N, 0, 0)]^T$. Since $g(x, 0, 0) \leq 0$ it follows that $\tilde{Q} \geq 0$. Next note by the mean value theorem that

$$\psi_i(u) = -h^2[g_u(x_i, \xi_i(u))u_i + g_{u'}(x_i, \xi_i(u))\frac{u_{i+1} - u_{i-1}}{2h}], \tag{2.15}$$

where $\xi_i(u)$ is on the line segment connecting $(0,0)$ and $(u_i, \frac{u_{i+1} - u_{i-1}}{2h})$. Thus we can write $\psi(u) = \Psi(u)u$ where Ψ is the tridiagonal matrix

$$\Psi(u) = -h^2 tri(\frac{-1}{2h} g_{u'}(x_i, \xi_i(u)), g_u(x_i, \xi_i(u)), \frac{-1}{2h} g_{u'}(x_i, \xi_i(u))). \tag{2.15}$$

For $h < 2/M$, (i) follows by defining $u_0 = [1, \ldots, 1]^T$, since $C - \Psi(u_0)$ is diagonally dominant, with strict dominance in at least one row (e.g., the first).

To show (ii), it is easily verified that $\psi' = \Psi(u)$ with ξ_i defined in (2.15) as

$$\xi_i(u) = (u_i, \frac{u_{i+1} - u_{i-1}}{2h}).$$

Thus $C + \psi'$ is irreducibly dominant; therefore $C + \psi'$ is invertible for all $u \geq 0$.

We use the representation $\psi(u) = \Psi(u)u$ to prove the coercivity condition (iii). Note in general (for $h < 2/M$) that $C + \Psi$ is tridiagonal,

$$C + \Psi(u) = tri(1 - \rho_i, -2, 1 + \rho_i),$$

with $g_u \geq 0$ and $0 \leq |\rho_i| \leq \mu < 1$. Let $\sigma_{min}(\rho)$ denote the minimum singular value of the matrix above where $\rho = [\rho_1, ..., \rho_N]$. Continuity of $\sigma_{min}(\cdot)$ together with the compactness of $\{\rho : |\rho| \leq \mu\}$

shows that the minimum of $o_{min}(\rho)$ is achieved. This minimum cannot be zero since $C + \Psi(u)$ is irreducibly dominant; hence invertible for all $u \geq 0$. Thus it follows that $|Cu + \Psi(u)u| \to \infty$ as $|u| \to \infty$.

To prove (iv) the representation $\psi(u) = \Psi(u)u$ once again. Suppose some component $u_i = 0$ and

$$Cu + \Psi(u)u + Q^* = 0, \quad \text{with} \quad Q^* \geq 0, \quad Q^* \neq 0.$$

The i^{th} equation from the system above has the form

$$(1 - \rho_i)u_{i-1} - (2 + h^2 g_u)u_i + (1 + \rho_i)u_{i+1} + Q_i^* = 0.$$

Since $|\rho_i| < 1$ and $u_{i-1}, u_{i+1} \geq 0$, it follows that if $u_i = 0$, then $u_{i-1}, u_{i+1}, Q_i^* = 0$. From this it follows that $u = 0$ and $Q^* = 0$, a contradiction. Therefore $u > 0$, proving (iv). Finally, it is also a straightforward matter to prove uniqueness of solutions. To see this, suppose u and w are two solutions, then we have

$$C(u - w) + \phi(u) - \phi(w) = 0.$$

Again using the mean value theorem, we can write

$$\phi(u) - \phi(w) = \Phi(u - w),$$

where Φ is a tridiagonal matrix of the form

$$\Phi = -h^2 tri(\frac{1}{2h} g_{u'}(x_i, \xi_i), g_u(x_i, \xi_i), \frac{-1}{2h} g_{u'}(x_i, \xi_i)),$$

with ξ_i lying on the segment connecting $(u_i, (u_{i+1} - u_{i-1})/2h)$ and $(w_i, (w_{i+1} - w_{i-1})/2h)$. Thus $C(u - w) + \Phi(u - w) = 0$. But again, $C + \Phi$ is irreducibly diagonally dominant, hence $u = w$.

We now return to equation (1.6) with temperature dependent conductances. A slight change in notation is made here as we write x_b for fixed boundary nodes, and x for the non boundary nodes. We write (1.5) in the form

$$C_{11}(x)x + R_{11}D_1(x) + C_{12}(x)(x_b) + R_{12}D_2(x_b) + Q = 0. \tag{2.16}$$

Theorem 2.4. Let C_{11} and C_{12} be C^2 functions and suppose $Q + C_{12}(x)(x_b) + R_{12}D_2(x_b) \geq 0$ for all $x \geq 0$, with at least one positive component. Suppose further that for each x, $C_{11}(x) + R$ is irreducibly dominant. Then if $\sup_x |C(x)| < \infty$ and $\inf_x o_{min}(C(x) + R) > 0$, (2.16) has a solution with $x \geq 0$.

Proof. Fix $x_0 > 0$ and consider the equation

$$C_{11}(x_0)x + R_{11}D_1(x) + C_{12}(x_0)(x_b) + R_{12}D_2(x_b) + Q = 0. \tag{2.17}$$

By Theorem 2.2 equation (2.17) has a unique solution for any Q with $Q + C_{12}(x_0)(x_b) + R_{12}D_2(x_b) \geq 0$ where at least one component is positivet. Now consider the homotopy $H(\lambda, x)$

$$H(\lambda, x) = [(1 - \lambda)C_{11}(x_0) + \lambda C_{11}(x)]x + R_{11}D(x) + [(1 - \lambda)C_{12}(x_0) + \lambda C_{12}(x)]x_b + R_{12}D(x_b) + Q. \tag{2.18}$$

A value y in the range space of H is called *regular* if the Jacobian of H has full rank at $H^{-1}(y)$. Saard's theorem [10] guarantees that almost every value (in the sense of Lebesgue measure) is

regular. Thus we can find a sequence $\{Q_N\}$ such that $Q_N \to Q$, $Q_N \geq Q$, and $Q - Q_N$ is a regular value of H for each N.

Now define H_N with domain $(0,1) \times R^N$ by

$$H_N(\lambda, x) = [(1-\lambda)C_{11}(x_0) + \lambda C_{11}(x)]x + R_{11}D(x) + [(1-\lambda)C_{12}(x_0) + \lambda C_{12}(x)]x_b + R_{12}D(x_b) + Q_N.$$

Zero is a regular value for H_N, and in particular $H_N^{-1}(0)$ is a one dimensional manifold. Hence, each component of $H_N^{-1}(0)$ is diffeomorphic either to a circle or an open interval $[1\ 0]$. The component whose closure contains $(0, x_0)$ is not diffeomorphic to a circle since

$$\frac{\partial H_N(0,x)}{\partial x}$$

is invertible, and by the implicit function theorem, solutions to $H_N(\lambda, x) \quad 0$ are unique in a neighborhood of $(0, x_0)$ for $|\lambda|$ sufficiently small. Also note that for each λ,

$$\begin{aligned}[(1-\lambda)C_{12}(x_0) + \lambda C_{12}(x)]x_b + R_{12}D(x_b) + Q_N = \ & (1-\lambda)[C_{12}(x_0)x_b + R_{12}D(x_b) + Q_N] \\ & \lambda[C_{12}(x)x_b + R_{12}D(x_b) + Q_N]\end{aligned}$$

Both terms on the right above are nonnegative since $Q_N \geq Q$, thus

$$[(1-\lambda)C_{12}(x_0) + \lambda C_{12}(x)]x_b + R_{12}D(x_b) + Q_N \geq 0,$$

with at least one positive component. Arguing as in the proof of Theorem 2.2, the component of $H_N(\lambda, x) = 0 \subset (0,1) \times R^n$ with limit point $(0, x_0)$ is contained in $(0,1) \times R^n_+$. Furthermore, since Theorem 2.2 shows that the solution to $H(0,x) \quad 0$ is unique, and that

$$[(1-\lambda)C_{11}(x_0) + \lambda C_{11}(x)]x + R_{11}D(x)$$

is coercive for each λ (this is proved in the same manner as in Lemma 2.1.), standard arguments [4,18] show that this component has a limit point $\{1\} \times x_N$. By continuity, $H_N(1, x_N) = 0$. And since it is evident that the sequence of solutions $\{x_N\}$ is bounded (from the coercivity assumption), it contains at least one limit point, i.e. there exists a subsequence x_{N_k} with $x_{N_k} \to x^*$. Then writing

$$\begin{aligned}H(1,x^*) &= H(1,x^*) - H_{N_k}(1,x^*) + H_{N_k}(1,x^*) - H_{N_k}(1,x_{N_k}) + H_{N_k}(1,x_{N_k}) \\ &= Q - Q_{N_k} + H_{N_k}(1,x^*) - H_{N_k}(1,x_{N_k}),\end{aligned}$$

it follows that $H(1,x^*) = 0$ since $Q_{N_k} \to Q$ and $x_{N_k} \to x^*$. And clearly $x^* \geq 0$ since $x_N > 0$ for every N. ///

3. Solution methods. For the systems defined in (1.5) and (2.11) (corresponding to Theorem 2.2 and Example 2.3), the homotopy method of proof suggests curve following algorithms as a numerical means for finding solutions. It was shown in Theorem 2.2 that there is a unique C^1 curve $x : [0,1] \to R^n$ with $H(x(\lambda), \lambda) = 0$ for all λ by choosing $x(0) = \rho[1 \cdots 1]^T$ with $\rho > 0$. For the homotopy defined in (2.1) this leads to the differential equation

$$x'(\lambda) \quad F'(x)^{-1}(Q - Q_0) \tag{3.1}$$

where we recall F' f]'0)]) (2.2) as

$$F'(x) = C + RD_x; \quad D_x = diag(4x_1^3, ..., 4x_n^3). \tag{3.2}$$

The solution to $F(x)$ () is then the solution to (3.1) at λ 1. A similar differential equation can be developed for the discretized boundary value problem in Example 2.3.

The corresponding differentials remain invertible along the solution paths so that there are no turning points, and the paths can be parameterized by λ. '1'1111s the ode of (3.1) can be solved or various other path following algorithms can be implemented [2,12,1 '1]. (We not e t hat these references also contain more general path following algorithms that deal with turning points and bifurcations.) Any of these methods gives rise to a globally convergent algorithm for t he problems (1 .6) and (2.12).

A path following technique of 1)euflhard [6] uses information along the path to generate bounds for which Newton's method is guaranteed to converge. A similar strategy can be employed here for solving (1.6) by moving along the curve $H(\lambda, x(\lambda)$ 0, whilst conforming to the *a priori* estimates required 1 ay the Newton Kantorovich (11(!01'(? n.

Thesecond derivative of F will be of interest in developing this idea since it figures prominently in the analysis of Newton's method. Wc have, noting that F'' is a bilinear mapping,

$$[F''(x)h] \quad (k) \quad 12R \begin{pmatrix} x_1^2 h_1 k_1 \\ \vdots \\ x_n^2 h_n k_n \end{pmatrix}. \tag{3.3}$$

For complet eness we state the Newton Kantoravich theorem (see for example [3]).

Newton- Kantorovich Theorem. Let x_0 be an initial estimate of the solution to (1 .2) and assum e that $F'(x_0)$ is an invertible matrix. 1 et β and η denote t wo constants such that

$$|F'(x_0)^{-1}| \leq \beta, \quad |F'(x_0)^{-1}(F(x_0))| \leq \eta, \tag{3.4}$$

and suppose there exists $r, K_r > 0$ with

$$|F''(x)| \leq K_r \quad \text{for all } x \text{ with } \quad |x - x_0| \leq r \tag{3.5}$$

Then if the constant $h = \beta\eta K_r$ satisfies

$$h \leq 1/2 \quad \text{and} \quad \frac{1}{h}[1 - \sqrt{1 - 2h}]\eta \leq r, \tag{3.6}$$

the Newton iteration

$$x_{k+1} = x_k - F'(x_k)^{-1}(F(x_k))$$

converges to a solution of $F(x)$ 0. Fu rthermore t his solut ion is unique in any closed ball centered at x_0 with radius less than r, and the convergence is quadr alit. / / /

Note that since $h \to 1/h[1 - \sqrt{12h}]$ is increasing on the int er val [(), 1 / 2], we may take $2\eta < 7$ for the second condition in (3.6).

Now suppose (x^*, λ^*) solves 11 (:/", λ^*) O for some $\lambda^* \in$ (O, 1), i.e.

$$Cx^* + RD(x^*) + Q + \lambda^* \quad (Q \quad Q_0) \; 0. \tag{3.7}$$

The next iterate is generated by solving

$$Cx + RD(x) + Q^* + t(Q \; Q^* \;) \; 0, \tag{3.8}$$

where $Q^* = Q + \lambda^*(Q - Q0)$, and $t > 0$ is CI1OSCI1 as t 11(! maximmmscalar for which the hypotheses of the Newton Kantoravich theorem can be verified. In gener al , at the k'" step the current solution is used to initialize a Newton iteration t o solve

$$Cx + RD(x) + Q_k = 0, \quad \text{where} \quad Q_k = Q_{k-1} + t_k(Q - Q_{k-1}). \tag{3.9}$$

W (: note that the sequence $\{Q_k\}$ generated in this manner lies on the segment connecting Q_0 to Q. 1 lence, if an iteration admitting the value $t_n = 1$ is obtained, the solution has been found. We will snow below how to generate $\{t_k\}$, and pr ove that t he iteration (3. 9) terminates after a finite number of steps.

To determine the value of t in (3.8) above define

$$\alpha = |F'(x*')^{-1}(Q - Q^*)|, \tag{3,"10}$$

and note that

$$\eta = |F'(x^*)^{-1}|Cx^* + R D(x^*) + Q^*| \leq t\alpha, \quad /(Q - Q^*)|| \tag{3.1 1}$$

since

$$Cx^* + RD(x^*) + Q' = 0.$$

Next, using the definition of F''' we have for $|x - x^*| < 7$,

$$\sup_x |F'''(x)|_\infty \leq 12|R|_\infty \max_i |x_i^* + r|^2. \tag{3.12}$$

Thus the hypotheses (3.4) (3.6) are satisfied so long as

$$t < \frac{r}{2\alpha}, \quad t < \frac{1}{2\alpha K_r \beta} \tag{3.13}$$

To maximize the value of t above, let $f_1(7") = r/2\alpha$, and $f_2(r) = 1/2\alpha K_r \beta$, with K_r given in (3.5). Define $f(r) = \min(f_1(r), f_2(r))$ and observe that for any 7', if $t < f(r)$, (3.6) is satisfied. Thus we maximize $f(r)$. Now f_1 is an increasing function of r and f_2 is a decreasing function of r on $[0, \infty)$. Furthermore, since $f_1(0) = 0$, and $f_2(0) > 0$, it follows that f is maximized when $f_1 = f_2$. This occurs at the smallest positive real solution t o t he cubic equat ion

$$r^3 + 2x_i^* r^2 + x_i^{*2} r - \frac{1}{12|R|_\infty \beta} = 0, \tag{3,14}$$

wl iere x_i^* denotes the maximum component of x^*. It is rosily verified that the cubic above has a single real solution, and this solution is positive. Once this solution is found, we set $t = r/2\alpha$.

Along the solution curve S, $x_i(\lambda)$ and β are both continuous. Because S is a compact set, x_i and β are both bounded with $\min(x_i^*) > 0$ and $\max(\beta) < \infty$. Thus the smallest positive real solution to this cubic over all points along the curve is bounded away from zero. Now as $Q^* \to Q$, it follows from (3.1O) that $\alpha \to 0$ since β is uniformly bounded on S. Hence, for Q^* sufficiently close to Q, say $|Q - Q^*| < \delta$, we have $t > 1$. To show that the iteration defined in (3.9) converges it suffices to show that after a finite number of steps an iterate will get within a δ neighborhood of Q. To see this, consider the di fference scheme

$$Q_{k+1} = Q_k + t_{k+1}(Q - Q_k)$$

and suppose $|Q - Q_k| > \delta$ for all k. Then necessarily $t_k \to 0$, which is impossible since $\{t_k\}$ is bounded away from zero. Hence for sufficiently large n, we get $|Q_n - Q| < \delta$, and so the algorithm terminates after a finite number of steps.

Of course efficient methods for path following can be developed for this problem since the curve is unique, there are no turning points, and as we are only interested in the value $x(1)$, it is not necessary to follow the curve exactly. Stepsize adjustment schemes based on how quickly the Newton corrector stage converges can be implemented, as well as the use of an Euler or other predictor steps since the differential at the current iterate is available [1,2,12].

For the discretized boundary value problem in Example 2.3, More ["14] used a generalization of diagonal dominance to nonlinear functions to prove that the nonlinear Gauss Seidel and Jacobi iterations are globally convergent. Similar results were obtained earlier by Rheinboldt using an extension of M matrices to nonlinear functions [15]. However, this particular extension required the domain of g in (2.11) to be defined on the entire real line, instead of the nonnegative axis, and convergence to a nonnegative solution was not established. Homotopy methods have been used previously for the discretized problem and also in conjunction with shooting methods for numerical solution of the boundary value problem [1,6,12,18,19].

The presence of the quartic term prohibits direct application of the methods in [14,'15] to establish global convergence of these algorithms for (1.6), although the nonlinear Gauss Seidel and Jacobi iterations are commonly used and perform well in practice. It is in fact straightforward to demonstrate that these iterations are at least locally convergent for (1.6).

To show this convergence for the nonlinear Jacobi iteration we define the function G as the diagonal of the mapping F

$$G(x) = \begin{pmatrix} c_{11}x_1 + r_{11}x_1^4 \\ \vdots \\ c_{nn}x_n + r_{nn}x_n^4 \end{pmatrix}.$$

Writing $F(x) = H(x) - G(x)$, where $H(x) = F(x) + G(x)$ we solve for the fixed point of the function ϕ,

$$\phi(x) = G^{-1}(H(x)), \tag{3.14}$$

which is equivalent to finding $G(x) = H(x)$ and hence, $F(x) = 0$. Assuming ϕ' is continuous, a sufficient condition for the convergence of the fixed point iteration in a neighborhood of the solution x^* is that $\rho(\phi'(x^*)) < 1$, where ρ denotes the spectral radius. From (3.14) we obtain

$$\phi' = (G^{-1})'(H(x))H'(x). \tag{3.15}$$

And by the inverse function theorem and the definition of H,

$$\begin{aligned} \phi' &= G'^{-1} H'(x) \\ &= I + G'^{-1} F'. \end{aligned} \tag{3.16}$$

Noting that

$$G'(x) = -diag(C + RD_x),$$

using the diagonal dominance of F' and irreducibility we will show that $\rho(I + G'^{-1}F') < 1$ holds everywhere, so that the nonlinear Jacobi iteration, $G(x^{N+1}) = H(x^N)$, is locally convergent. To see this, first observe that $G'^{-T}F'^T$ is irreducibly dominant with negative ones on the diagonal. Hence, every eigenvalue of $G'^{-1T}F'^T$ has negative real part, and because of the irreducible dominance, $\rho(G'^{-T}F'^T) < |G'^{-T}F'^T|_\infty \leq 2$. Thus, $\rho(I + G'^{-T}F'^T) < 1$, verifying local convergence. Local convergence for the Gauss Seidel iterations follow since the associated spectral radius for the

Gauss Seidel update is strictly less than for the Jacobi iteration whenever the Jacobi iteration is convergent ([3]).

Another class of globally convergent algorithms for (1.6) and (2.12) follow from the observation that the nonlinear least squares problem,

$$\min_{x>0} J(x) \equiv |F(x)|^2,$$

has a unique stationary point for $x > 0$ corresponding to $F(x) = 0$ since it has already been established in Theorem 2.2 and Example 2.3 that $F(x) = 0$ has a unique solution for $x > 0$ and also that F' is invertible for $x > 0$. Thus the class of bound constraint algorithms represent an alternative to the solution methods already discussed. For example, trust region methods enjoy, under mild assumptions, global convergence to a Kuhn Tucker point of the inequality constrained problem above [8]. In fact if C in (1.6) is invertible, we can show that the solution $F(x) = 0$ is the unique Kuhn Tucker point, thus making the trust region algorithm globally convergent to the solution $F(x) = 0$. This is easily demonstrated: The Kuhn Tucker conditions for this problem are

$$F'^T(F(x)) - \lambda = 0,$$

$$x, \lambda \geq 0$$

$$\lambda_i x_i = 0, \quad \text{for all } i$$

where λ denotes the Lagrange multiplier corresponding to the inequality constraint $x \geq 0$. Since F'^T is invertible, the first condition is equivalent to

$$F(x) - F'^{-T}\lambda = 0.$$

However, since F'^T is irreducibly dominant with positive entries off the diagonal, $-F'^{-T} \geq 0$ [16] (i.e., all of its entries are nonnegative). And since $\lambda \geq 0$, $F'^{-T}\lambda \geq 0$. Suppose some constraint is active, say $x_i = 0$. By the arguments in Theorem 2.2 it follows that $x = 0$, since $Q - F'^{-T}\lambda \geq 0$. This is impossible. Thus, the constrained optimization problem has only the single KT point at $F(x) = 0$. We note that this argument works equally well for Example 2.3.

The situation for the variable conductor problem is somewhat different with respect to the implementation of numerical algorithms. The Newton path following algorithm for nonvariable conductors was based on the invertibility of $\partial H(\lambda, x)/\partial x$ along the solution path. Theorem 2.4 does not establish this result, and in general it is possible for the path not to be monotonic in λ. However, several path following algorithms exist for this more general situation [1,2,11,12,17].

The noninvertibility of $\partial H(\lambda, x)/\partial x$, if it occurs, has a physical interpretation. At $\lambda = 0$ the eigenvalues of $\partial H(\lambda, x)/\partial x$ in (2.17) Since the degree is a homotopic invarian, $deg(H(\lambda, x)) = deg(H(0, x))$ for $\lambda \in [0, 1]$ [10]. If λ ceases to be monotonic, and the curve folds backwards, the number of solutions to $H(\lambda, x) = 0$ is greater than one. Thus the determinant must change sign. This implies that at least one eigenvalue passes through zero and becomes positive. This in turn implies that the solutions along this part of the curve are unstable in the sense that the steady state solution of the associated differential equation

$$\frac{dx}{dt} = H(\lambda, x(t)), \quad x(0) = x_0 \quad \text{where} \quad H(\lambda, x_0) = 0,$$

is not stable.

Acknowledgement

This publication was prepared by the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

References

[1] E. Allgower and K. Georg, Simplicial and continuation methods for approximating fixed points and solutions to systems of equations, SIAM Review, 22, January, 1980, pp. 28–85.

[2] E. Allgower and K. Georg, *Numerical Continuation Methods: An Introduction,* Springer-Verlag, New York, 1990.

[3] E. K. Blum, *Numerical Analysis and Computation: Theory and Practice*, Addison Wesley Publishing Company, Reading, MA, 1972.

[4] S. N. Chow, J. Mallet-Paret and J. A. Yorke, Finding zeros of maps: homotopy methods that are constructive with probability one, Math. Comp., 32, 1978, pp. 887–899.

[5] B. A. Cullimore, SINDA '85/FLUINT System Improved Numerical Differencing Analyzer and Fluid Integrator, Version 2.3, Martin Marietta, 1990.

[6] P. Deuflhard, A stepsize control for continuation methods and its special application to multiple shooting techniques, Numer. Math., 33, 1979, pp. 115–146.

[7] J. Fanson, Ed., *Space Infrared Telescope Facility Baseline Observatory Design for a Delta Launch*, JPL D-12375, Rev A.

[8] R. Fletcher, *Practical Methods of Optimization, Second Edition*, John Wiley and Sons, New York, 1987.

[9] D. G. Gilmore, ed., Satellite Thermal Control Handbook, The Aerospace Corporation Press, El Segundo, 1994.

[10] V. Guillemin and A. Pollack, *Differential Topology*, Prentice-Hall, Inc., Englewood Cliffs, NJ, 1974.

[11] R. B. Kearfott and Z. Xing, An interval step control for continuation methods, SIAM J. Numer. Anal., 31, June 1994, pp. 892–914.

[12] H. B. Keller, *Lectures on Numerical Methods in Bifurcation Problems*, Springer-Verlag, Berlin, 1987.

[13] C. K. Krisnaprakas, Application of accelerated iterative methods for solution of thermal models of spacecraft, J. Spacecraft and Rockets, 32, July, 1995, pp. 608–611.

[14] J. J. More, Nonlinear generalizations of matrix diagonal dominance with application to Gauss-Seidel iterations, SIAM J. Numer. Anal., 9, June 1972, pp. 357–378.

[15] W. C. Rheinboldt, On *M*-Functions and their application to nonlinear Gauss-Seidel iterations and to network flows, J. Math. Anal. Appl., 32, 1970, pp. 274–307.

[16] R. Varga, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, N. J., 1962.

[17] L. T. Watson, Numerical linear algebra aspects of globally convergent homotopy methods, SIAM Rev., 28, December 1986, pp. 529–545.

[18] L. T. Watson, An algorithm that is globally convergent with probability one for a class of nonlinear two point boundary value problems, SIAM J. Numer. Anal., 16, June 1979, pp. 394–401.

[19] L. T. Watson, solving finite difference approximations to nonlinear two-point boundary value problems by a homotopy method, SIAM J. Sci. Stat. Comput. 1, 16, December 1980, pp. 467–480.

[20] O. C. Zankiewicz, Finite Element methods in thermal problems, in *Numerical Methods in Heat Transfer*, R. W. Lewis, K. Morgan, and O.C. Zankiewicz, Ed., pp. 1- 25 J. Wiley and Sons, Bristol, England, 1981.